

# Hydrophobicity and Unique Folding of Selected Polymers

Michele Vendruscolo

*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*

In suitable environments, proteins, nucleic acids and certain synthetic polymers fold into unique conformations. This work shows that it is possible to construct lattice models of foldable heteropolymers by expressing the energy only in terms of individual properties of monomers, such as the exposure to the solvent and the steric factor.

PACS numbers: 87.15.By, 87.10.+e

It is generally believed that the hydrophobic interaction plays a major role in protein folding [1–10]. Under physiological conditions, non polar amino acids are buried inside the core of the native state of a protein to avoid contact with water molecules. A long standing question is to what extent other non-covalent forces, such as hydrogen bonding, electrostatic and van der Waals interactions contribute to stabilize the folded state [1–4].

Unraveling the different roles played by these interactions will have a considerable impact in different areas of research in biophysics, such as the prediction of protein structures [1,2,5–7] the design of synthetic drugs [6,11–13], and the production of self-assembling non-biological polymers [14] and other polymeric materials [15].

With the advent of genome projects [16] a wide gap is opening between the number of known protein sequences and their correspondent structures [17]. The bottleneck in protein structure prediction is at present largely due to the incorrect treatment of the interactions [18–20]. The current state of the art is highlighted by a recent study by Fisher and Eisenberg [21]. They carried out the assignment of structures to the sequences encoded in the complete genome of *Mycoplasma genitalium*. Among the complete set of 468 sequences, they were able to assign 103 (22%) of them to a structure with high confidence. They used homology modelling and threading techniques that are at present the most successful prediction tools available. Homology modelling is based on the observation that the exact identity of amino acids is not crucial for maintaining the overall fold of a protein [22]. Proteins that differ as much as 70% in their sequences usually share the same fold. Thus, it is possible to predict the conformation of a sequence by using a set of experimentally determined protein structures with similar amino acid composition. Threading relies on the surprising fact that more distantly related proteins, whose sequence sim-

ilarity is close to the threshold of pure randomness, do sometimes share the same fold [4]. The search for *compatibility* between sequence and structure has inspired various techniques to single out the native state of a protein from a library of alternative structures [4,22,23]. The screening is typically carried out by assigning an energy-like function that incorporates the compatibility of each amino acid to its local environment. Compatibility is described in terms of charge, polarity and secondary structure content, within a given conformation. Details in the local environment play a crucial role also in RNA folding. A key ingredient in this case is given by metal ion coordination numbers [24]. Likewise and rather surprisingly, a non-biological polymer (an aromatic hydrocarbon) has been recently designed which is able to fold into a unique helical structure having a large cavity, supposedly under the effect of the hydrophobic interaction [14].

This letter contributes to the development of a rational treatment of the hydrophobic interaction at the single monomer level. We show that it is possible to construct minimalistic lattice models of foldable heteropolymers, by introducing an energy function that depends only on individual residues' environments. A model will be called "foldable" if there are sequences, either randomly chosen or selected, with a unique, thermodynamically stable and kinetically reachable ground state [2,3,6–9,20,25–27]. We adopt a simple approximation for the energy which accounts both for the propensity to be exposed to the solvent and for the excluded volume effects due to the different sizes of the monomers. Although naturally existing or synthesized polymers, such as proteins, nucleic acids and tailored hydrocarbons, are characterized by much more complex interactions, the main focus here is on the fact that the unifying feature is the tendency to avoid contact with the solvent by some species of monomers. Previous theoretical studies concentrated mainly on the treatment

of pairwise contact energies [2,18,20,28–32]. This is in contrast with the present study, in which the hydrophobic effect is investigated at the individual particle level.

Lattice models, although often criticized [33], have been recognized to capture some of the most relevant thermodynamic features of the folding process [34], such as the existence of a unique ground state, amenable to exact computations, and the cooperativity of the transition. Even key dynamical processes, such as the nucleation-condensation mechanism [35], have been validated with the help of lattice models [36].

Results obtained on a 2D square lattice are presented first. On a lattice, a polymer is represented as a connected chain of  $N$  monomers. Hydrophobicity and steric factors can be modeled as the tendency of a monomer to have a specific number of non-bonded nearest-neighbors. We define the hydrophobic model  $\text{HM}_1$  by expressing the energy  $E_1$  as,

$$E_1 = \sum_{i=1}^N |n(a_i) - \bar{n}(a_i)|, \quad (1)$$

where  $n(a_i)$  is the number of non-bonded nearest-neighbors of monomer of species  $a_i$  in position  $i$  along the chain, and  $\bar{n}(a_i)$  is the ideal value of  $n(a_i)$ . This expression was first proposed by Hao and Scheraga, who supplemented it to a pairwise energy term [37]. They presented a method to optimize energy parameters to obtain lattice models of foldable polymers. Other previous work has been devoted to the hydrophobic interaction, although without specifically disentangling it from other interactions. Mirny and Domany [32] introduced explicitly an hydrophobic term in the energy function and they performed various tests of fold recognition and dynamics. In a recent work Li, Tang and Wingreen [38] discussed the “designability principle” [9] in terms of a “binary” model with two species of amino acids, where the energy is expressed in terms of the exposure to the solvent only. The model proposed here is much more general and no major modification is required to extend it to the treatment of realistic models of foldable heteropolymers, as for example, in a “contact map” representation of protein structure [32,39].

We will first investigate if in 2D the  $\text{HM}_1$  model gives rise to foldable sequences and we will compare our results with those obtained using the standard HP model [2]. In 2D, with the aid of the HP model [2,8,9,13], it has been demonstrated that it suffices to assume only 2 species of monomers to guarantee

uniqueness of the ground state, although probably not the right order of the folding transition [7]. In the HP model the energy is written in the pairwise contact approximation

$$E_{\text{pair}} = \sum_{j>i} U(a_i, a_j) \Delta_{ij}, \quad (2)$$

where  $a_i$  can be either  $H$  (hydrophobic) or  $P$  (polar) and  $\Delta_{ij}$  is a contact matrix, which is defined to be 1 if two monomers are non-bonded nearest-neighbor and 0 otherwise. The typical values for the interaction parameters are  $U(H, H) = -1$  and  $U(H, P) = U(P, P) = 0$  [2]. A chain of  $N = 16$  monomers is amenable to complete enumeration of all 802075 possible symmetry-unrelated conformations, either compact or not [2,12,13]. For the above mentioned choice of contact energy parameters, there are 1539 (2%) sequences among the  $2^{16} = 65536$  possible ones which have a unique ground state [12,13]. We compare this result with those obtained by using Eq.(1), setting  $\bar{n}(1) = 1.5$  (hydrophobic-like) and  $\bar{n}(2) = 0.4$  (polar-like). A larger number, 10178 (16%), of sequences was found to have a unique ground state.

We have also explored the case of three species of monomers, a number which, within a contact approximation of the interactions (as in Eq.(2)), is believed to epitomize the essential features of the interplay between folding and glass transitions in random heteropolymers [40]. We chose at random 20177 sequences among the 2018016 possible ones with fixed composition  $N(1) = 6$ ,  $N(2) = N(3) = 5$ , where  $N(a)$  is the number of monomers of species  $a$ . Choosing energy parameters  $\bar{n}(1) = 0.4$ ,  $\bar{n}(2) = 1.1$  and  $\bar{n}(3) = 1.8$ , 9439 (47%) sequences were found with a unique ground state.

In the spirit of Mirny and Domany [32], a more realistic form for the hydrophobic energy is given by

$$E_2 = \sum_{i=1}^N \beta(a_i) [n(a_i) - \bar{n}(a_i)]^2. \quad (3)$$

This expression will be referred to as hydrophobic model  $\text{HM}_2$ . The parameters  $\beta(a_i)$  capture the various degrees with which the different species of monomers tend to attain the preferred number  $\bar{n}(a_i)$  of contacts. In the case of 2 species of monomers, we repeated the same calculation as for the  $\text{HM}_1$  model. Letting here  $\beta(1) = \beta(2) = 1$ , we found 9821 (14%) sequences with a unique ground state. We observe that, at least from the above calculations in 2D, the approximation of the hydrophobic interaction proposed in this work is capable of yielding foldable sequences.

We now turn to the calculations in 3D which represent the essential part of this work, and further illustrate the extent to which the present model embodies foldability. It is known that the HP model is pathological in 3D, since it is rather uncommon to have a sequence with a unique ground state with a large gap above it [7,41], although the situation can be different with a choice of parameters favoring more collapsed structures [9,20]. We discuss here the general case of 20 species of monomers in the  $HM_2$  model. We compare these results with those obtained by using a common parametrization of the pairwise contact interaction matrix  $U(a_i, a_j)$  in Eq.(2), due to Miyazawa and Jernigan (MJ) [28], although other choices would be possible [28–32]. For the  $HM_2$  model, we derived the 40 parameters  $\bar{\pi}(a)$  and  $\beta(a)$  from a statistical analysis of the non redundant set of 246 protein structures reported by Hinds and Levitt [30]. The procedure, similar to that of Mirny and Domany [32], is straightforward. For each amino acid species  $a$ , we computed the average,  $\bar{\pi}(a)$ , and the standard deviation,  $\beta(a)$ , of the number of contacts it forms in the set of experimentally known crystal structures (see Table I). Two amino acids are said to be in contact if their  $C_\alpha$  atoms are closer than 8.5 Å in the native structure [39].

On the cubic lattice, the 103346 symmetry-unrelated maximally compact conformations of a polymer of length  $N = 27$  can be enumerated in a manageable computer time [9,43]. If it is guaranteed that the ground state is maximally compact, exact enumeration can be used to demonstrate its uniqueness. We adapted the energy parameters to the cubic lattice by matching the average number of contacts that a monomer forms on the  $3 \times 3 \times 3$  cube with the average of the ideal number of contacts,  $(1/20) \sum_{a=1,20} \bar{\pi}(a)$ . This result is obtained by rescaling the energy parameters in Table I by a factor 3.315.

To characterize foldability, we first investigate the thermodynamic stability of the ground states of random sequences. A typical measure of thermodynamic stability is given by the  $Z$  score [20,31], which is defined by  $Z = (E_n - \langle E \rangle) / \sigma$ , where  $E_n$  is the energy of the ground state,  $\langle E \rangle$  is the average energy, and  $\sigma$  the standard deviation in the distribution of the energy around the average. We measure the distribution of the  $Z$  scores for a set of 1000 random  $HM_2$  sequences. We found that only 2% of them had a unique lowest energy state (the “ground state” among maximally compact conformations). Moreover, on average the degeneracy was 22. For com-

parison, 99% of the 1000 random MJ sequences that we considered had a non degenerate lowest energy state, and the remaining ones had a very small degeneracy. The result of the comparison of the  $Z$  scores is shown in Fig. 1.

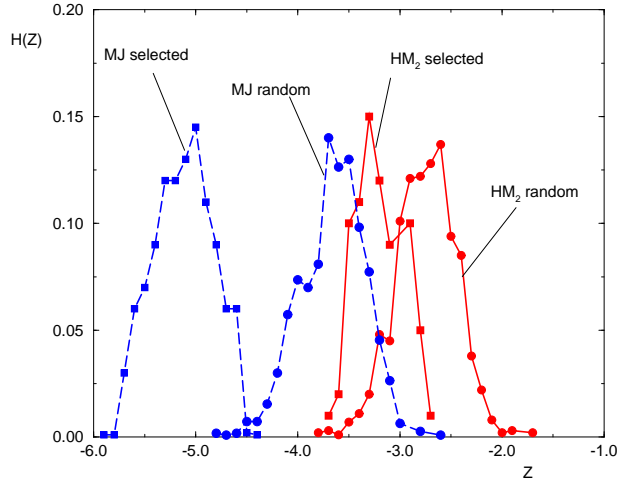


FIG. 1. Normalized histograms of the  $Z$  scores for the  $HM_2$  (full lines) and MJ (dotted lines) models. Circles refer to random sequences and squares to designed ones.

The low values of the  $Z$  score and the large degeneracy found for the  $HM_2$  case mark a shortcoming of enumerating only maximally compact conformations. By using other simulation techniques, such as the standard lattice Monte Carlo (SMC) [25,42], and the prune-enriched Rosenbluth method (PERM) [44], we easily found non-compact lower energy states for most of the considered  $HM_2$  sequences. It is known, however, that foldability is a property of *selected* sequences [5–7,10–13,26,27]. A way to demonstrate that the  $HM_2$  model is foldable is to show that it is possible to select sequences whose ground states are both unique *and* maximally compact. The usual design procedure [11], introduced to study pairwise interactions, prescribes to choose a target conformation and then to search in sequence space for the sequence with minimal energy onto such conformations. This procedure delivers a better  $Z$  score for the 1000 designed MJ sequences that we considered, as can be seen from Fig. 1. However, in the case of the  $HM_2$  model, we found that such technique is not sufficiently effective in designing out alternative conformations. A sequence design procedure similar to those proposed in Refs. [12,13] proved to be more effective. Sequences selected in this way

were found to have a unique ground state by exact enumeration among maximally compact conformations. More crucially, in no cases we have been able to reach lower energy states using dynamical simulation techniques such as the SMC and the PERM algorithms. The histogram of the  $Z$  score of the 100 HM<sub>2</sub> sequences selected in this way is shown in Fig. 1.

In summary, we have shown both in 2D and in 3D that it is possible to construct simple models of foldable heteropolymers by expressing the hydrophobic and the steric interactions at the level of individual monomers.

It is a pleasure to thank E. Domany and P. Grassberger for discussions.

- 
- [1] Creighton, T. E. (ed.), *Protein folding*, W. H. Freeman and Company, New York. (1992).
  - [2] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, *Protein Sci.* **3**, 561 (1995).
  - [3] J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).
  - [4] D. T. Jones, and J. M. Thornton, *Curr. Opin. Struct. Biol.* **6**, 210 (1996).
  - [5] T. Garel, H. Orland, and D. Thirumalai, in *New Developments in Theoretical Studies of Proteins* R. Elber (ed.), World Scientific, Singapore. (1996).
  - [6] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Rev. Mod. Phys.*, in press.
  - [7] E. I. Shakhnovich, *Curr. Opin. Struct. Biol.* **7**, 29 (1997).
  - [8] C. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **90**, 6369 (1993).
  - [9] H. Li, R. Hellig, C. Tang, and N. Wingreen, *Science* **273**, 666 (1996).
  - [10] A. Irbäck, C. Peterson, and F. Potthast, *Proc. Natl. Acad. Sci. USA* **93**, 9533 (1996).
  - [11] E. I. Shakhnovich and A. M. Gutin, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7195 (1993).
  - [12] J. M. Deutsch and T. Kurosky, *Phys. Rev. Lett.* **76**, 323 (1996).
  - [13] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, *Phys. Rev. Lett.* **77**, 1901 (1996).
  - [14] J. C. Nelson, J. G. Saven, J. S. Moore, and P. G. Wolynes, *Science* **277**, 1793 (1997).
  - [15] M. Muthukumar, C. K. Ober, and E. L. Thomas, *Science* **277**, 1225 (1997).
  - [16] D. Duboule, *Science* **278**, 555 (1997), and references therein.
  - [17] J. L. Sussman, *Nature Struct. Biol.* **4**, 517 (1997).
  - [18] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *J. Chem. Phys.* **103**, 9482 (1995).
  - [19] A. Finkelstein, *Curr. Opin. Struct. Biol.* **7**, 60 (1997).
  - [20] M. Vendruscolo, A. Maritan, and J. R. Banavar, *Phys. Rev. Lett.* **78**, 3967 (1997).
  - [21] D. Fisher, and D. Eisenberg, *Proc. Natl. Acad. Sci. USA* **94**, 11929 (1997).
  - [22] D. Fisher, D. Rice, J. U. Bowie, and D. Eisenberg, *FASEB J.* **10**, 126 (1996).
  - [23] J. U. Bowie, R. Luethy, and D. Eisenberg, *Science* **253**, 164 (1991).
  - [24] J. A. Doudna and E. A. Doherty, *Fold. Des.* **2**, R65 (1997).
  - [25] A. Sali, E. I. Shakhnovich, and M. Karplus, *Nature* **369**, 248 (1994).
  - [26] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins* **21**, 167 (1995).
  - [27] T. Veitshans, D. K. Klimov, and D. Thirumalai, *Fold. Des.* **2**, 1 (1997).
  - [28] S. Miyazawa and R. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
  - [29] R. A. Goldstein, Z. A. Luthey-Shulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **89**, 4918 (1992).
  - [30] D. A. Hinds, and M. Levitt, *J. Mol. Biol.* **243**, 668 (1994).
  - [31] L. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **264**, 1164 (1996).
  - [32] L. Mirny, and E. Domany, *Proteins* **26**, 391 (1996).
  - [33] B. Honig, and F. Cohen, *Fold. Des.* **1**, R17 (1996).
  - [34] E. I. Shakhnovich, *Fold. Des.* **1**, R50 (1996).
  - [35] A. R. Fersht, *Proc. Natl. Acad. Sci. USA* **92**, 10869 (1995).
  - [36] E. I. Shakhnovich, V. I. Abkevich, and O. Ptitsyn, *Nature* **379**, 96 (1996).
  - [37] M. H. Hao, and H. A. Scheraga, *Physica A* **244**, 124 (1997).
  - [38] H. Li, C. Tang, and N. Wingreen, preprint, cond-mat/9709083.
  - [39] M. Vendruscolo, E. Kussell, and E. Domany, *Fold. Des.* **2**, 295 (1997).
  - [40] J. N. Onuchic, N. D. Socci, Z. Luthey-Shulten and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **92** (1995) 3626.
  - [41] K. Yue, K. Fiebig, P. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 325 (1995).
  - [42] N. D. Socci and J. N. Onuchic, *J. Chem. Phys.* **103**, 4732 (1995).
  - [43] E. I. Shakhnovich and A. Gutin, *J. Chem. Phys.* **93**, 5967 (1990).
  - [44] P. Grassberger, *Phys. Rev. E* **56**, 3682 (1997).

ALA	GLU	GLN	ASP	ASN	LEU	GLY	LYS	SER	VAL	ARG	THR	PRO	ILE	MET	PHE	TYR	CYS	TRP	HIS
7.56	5.62	6.23	5.51	6.02	7.63	5.55	5.86	6.31	8.29	6.58	6.73	5.73	8.07	7.72	7.58	7.45	8.81	7.67	6.59
2.98	2.17	2.36	2.41	2.58	2.25	3.51	2.08	3.07	2.53	2.33	2.77	2.85	2.35	2.37	2.31	2.51	2.42	1.36	2.43

TABLE I. Mean  $\overline{n}(a)$  and standard deviation  $\beta(a)$  of the number of contacts of amino acids, obtained from a statistical analysis of a non redundant set of 246 protein structures.